

Statistik

- Grundlagen der deskriptiven Statistik
- Tabellarische und graphische Darstellung
- Lageparameter
- Streuungsparameter
- Zusammenhangsmaße
- Prognosemodelle

Grundlagen der deskriptiven Statistik

Grundbegriffe

Merkmalsträger - Objekte, über die Daten vorliegen

Grundgesamtheit - Alle potentiellen Merkmalsträger

Merkmal - Beobachtete Eigenschaft

Merkmalsausprägung - Möglicher Wert

Wertebereich - Gesamtheit aller realisierten Werte

Merkmalsausprägungen

Nominal - Kann man nicht ordnen

Ordinal - Rangfolge

Metrisch - Maßeinheit

Diskret - Nur bestimmte Ausprägungen

Stetig - Alle möglichen innerhalb eines Intervalls

Metrische Daten können **intervallskaliert** wie bei Grad Celsius oder **verhältnisskaliert** wie bei Grad Kelvin sein.

Besondere Vorsicht ist geboten, wenn nominale Werte **codiert** werden oder wenn stetige Merkmale künstlich **diskretisiert** werden.

Tabellarische und graphische Darstellung

Man kann für ein Merkmal die absolute Häufigkeit und die relative Häufigkeit (%) messen. Die absolute Häufigkeit liefert dabei auch die Größe der Stichprobe. Mit den relativen Häufigkeiten lassen sich besser Vergleiche anstellen.

Stetige Merkmale sollten zu Klassen zusammengefasst werden.

Absolute Häufigkeit

	M	W	Summe
WI	23	7	30
BW	12	11	23
Summe	35	18	53

Relative Häufigkeit

	M	W	Summe
WI	0,43	0,13	0,57
BW	0,23	0,21	0,43
Summe	0,66	0,34	1

Randverteilung

Bedingte Häufigkeit

→ Ein Merkmal fixieren

	M	W	Summe
WI	0,77	0,23	1
BW	0,52	0,48	1

Wenn die bedingten Merkmalsausprägungen mit den Randverteilungen übereinstimmen, sind sie unabhängig.

Die Unabhängigkeit wird eigentlich nie vollständig erreicht.

Lageparameter

Arithmetisches Mittel

$$\bar{x} = \frac{\text{Summe aller Werte}}{\text{Anzahl}}$$

→ Sehr Ausreißer anfällig

Median

Wert, der Reihe in zwei Teile zerlegt

$x_1 \ x_2 \ x_3 \ x_4 \ x_5$

$x_1 \ x_2 \ x_3 \ x_4 \ \rightarrow \ (x_2 + x_3) : 2$

→ relativ unempfindlich gegenüber Ausreißern

Modus

Häufigster Wert

→ Ausreißer robust

Nominal: Modus

Ordinal: Modus, Median

Metrisch: Modus, Median, Mittelwert

Perzentile

→ Schneidet Werte in zwei Teile

50-50 → Median

25-75 → Unteres Quartil

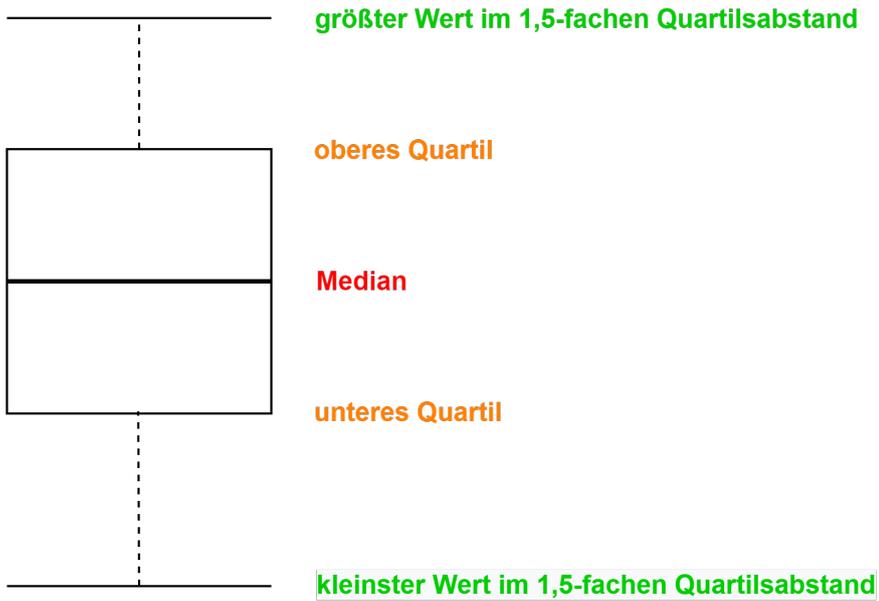
75-25 → Oberes Quartil

Der Boxplot

→ Graphische Darstellung der Verteilung

Ausreißer

Ausreißer



Ausreißer

Das geometrische Mittel

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \dots x_n}$$

Mit x_1 = Wachstumsfaktor, z.B. 1,05 bei 5% Zunahme oder 0,95 bei Abnahme

Streuungsparameter

Die Streuung der Werte gibt die Abweichungen vom Mittelwert an.
Für nominal skalierte Werte gibt es kein Streuungsmaß.

Spannweite

$$S = \text{größter Wert} - \text{kleinster Wert}$$

→ extrem empfindlich bei Ausreißern

Quartilsabstand

$$Q = \text{Oberes Quartil} - \text{Unteres Quartil}$$

Mittlere absolute Abweichung

$$M = \frac{\text{Summe (Beobachtungswert} - \text{Mittelwert)}}{\text{Anzahl}}$$

→ kein Vergleich verschiedener Merkmale möglich

Varianz

$$s^2 = \frac{\text{Summe (Beobachtungswert} - \text{Mittelwert})^2}{\text{Anzahl}}$$

→ schwer zu interpretieren
→ kein Vergleich verschiedener Merkmale

Standardabweichung

$$s = \sqrt{s^2} \quad \text{mit } s^2 = \text{Varianz}$$

→ gut zu interpretieren
→ kein Vergleich verschiedener Merkmale

Variationskoeffizient

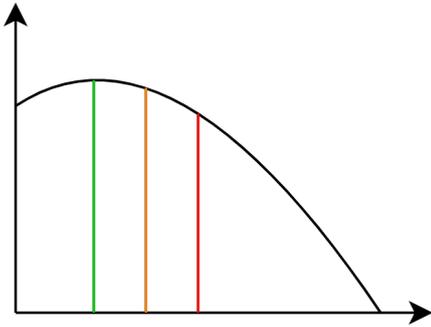
$$v = \frac{s}{\bar{x}} \quad \text{mit } s = \text{Standardabweichung und } \bar{x} = \text{Mittelwert}$$

→ relatives Maß

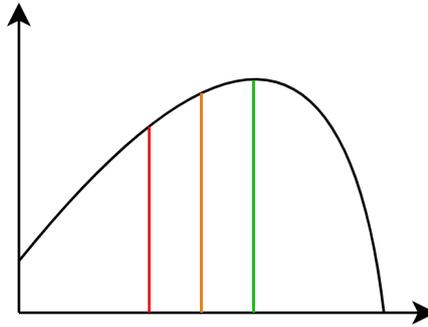
→ Vergleich verschiedener Merkmale möglich

Schiefte

rechts



links



Modus
Mittelwert
Median

Zusammenhangsmaße

→ Stärke und Richtung des statistischen Zusammenhangs

Nominale Daten

Chi Quadrat

Relative Häufigkeit

	M	W	Summe
WI	0,43	0,13	0,57
BW	0,23	0,21	0,43
Summe	0,66	0,34	1

Erwartete relative Häufigkeit

	M	W	Summe
WI	0,37	0,19	0,56
BW	0,29	0,15	0,44
Summe	0,66	0,34	1

$$0,66 * 0,56 = 0,37$$

$$\text{Chi - Quadrat} = N \cdot \left[\frac{(0,43 - 0,37)^2}{0,37} + \frac{(0,23 - 0,29)^2}{0,29} \right]$$

Cramers V

$$v = \sqrt{\frac{\text{Chi - Quadrat}}{N(\min(s, t) - 1)}}$$

→ $N \cdot (\text{höchste Anzahl Merkmalsausprägung} - 1)$

$$v = \sqrt{\frac{3,4291}{53(2 - 1)}} = 0,2542$$

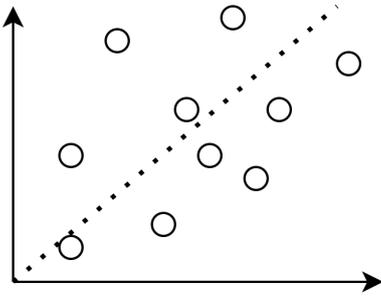
Bis 0,2 schwach
Bis 0,6 mittel
Bis 1,0 stark

Metrische Daten

Streudiagramm

→ graphische Darstellung

→ Je näher einer Gerade, desto stärker



Kovarianz

$$S_{xy} = \frac{(x - \bar{x})(y - \bar{y})}{N}$$

→ nicht gut interpretierbar

Korrelationskoeffizient

$$r = \frac{S_{xy}}{S_x S_y} \quad \text{mit } S_{xy} = \text{Kovarianz}; S_x = \text{Standardabweichung}$$

- $-1 \leq r \leq -0,6$ Starker negativer Zusammenhang
- $-0,6 < r \leq -0,2$ Mittlerer negativer Zusammenhang
- $-0,2 < r < 0$ Schwacher negativer Zusammenhang
- $0 < r \leq 0,2$ Schwacher positiver Zusammenhang
- $0,2 < r \leq 0,6$ Mittlerer positiver Zusammenhang
- $0,6 < r \leq 1$ Starker positiver Zusammenhang

Ordinale Daten

Rangkorrelationskoeffizient nach Spearman

Zunächst werden Rangzahlen verteilt. Bei mehreren gleichen Ausprägungen bekommen diese Daten das arithmetische Mittel der Ränge.

$$r_{SP} = \frac{\frac{1}{N} \sum_{i=1}^N (u_i - \bar{u})(v_i - \bar{v})}{S_u S_v}$$

Wobei $u = \text{Rang } x$ und $v = \text{Rang } y$

Prognosemodelle

Die Analyse statistischer Daten dient zur Vorhersage beobachteter Tatbestände.
Der vom Modell erklärte Teil sollte dabei anteilig groß sein.

$$y = b_1 x + b_2$$

$$b_1 = \frac{S_{xy}}{S_x^2} \quad \text{Mit } S_{xy} = \text{Kovarianz und } S_x^2 = \text{Varianz } (x)$$

$$b_2 = \bar{y} - b_1 \bar{x}$$

Das Bestimmungsmaß B gibt an, ob eine Prognose verwendet werden sollte, je näher an 1 desto besser (Anteil Varianz)

$$B = r^2 \quad \text{mit } r = \text{Korrelationskoeffizient}$$